DOCUMENT RESUME

ED 310 123 TM 013 696

AUTHOR Kogut, Jan
TITLE Detecting Aberrant Response Patterns in the Rasch Model. Rapport 87-3.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE 87
NOTE 45p.; Also cited as Project Psychometrische Aspecten van Item Banking No. 18.
AVAILABLE FROM Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Assisted Testing; Computer Simulation; Difficulty Level; Estimation (Mathematics); *Guessing (Tests); *Latent Trait Theory; Mathematical Models; *Statistical Bias; *Testing Problems
IDENTIFIERS Aberrance; Item Parameters; Iterative Methods; Person Fit Measures; *Rasch Model; *Response Patterns; Three Parameter Model

ABSTRACT

In this paper, the detection of response patterns aberrant from the Rasch model is considered. For this purpose, a new person fit index, recently developed by I. W. Molenaar (1987) and an iterative estimation procedure are used in a simulation study of Rasch model data mixed with aberrant data. Three kinds of aberrant response behavior are considered: (1) guessing to complete the test; (2) guessing in accordance with the three-parameter logistic model; and (3) responding with different abilities on different subsets of items. The power in detecting such aberrants is evaluated in two cases: when item difficulties are known; and when item difficulties are estimated from the data, including aberrants. The results reveal that, in the latter case, the estimates of the model parameters are biased and that the power of the index, as a consequence, is reduced. It is shown that, by using an iterative procedure, the recovery of the power of the index to the level obtained by known item difficulties is achieved. Furthermore, depending on the type of aberrance, a considerable reduction of the bias in the model parameters is possible. Finally, it is confirmed that this new index allows detection of aberrant response patterns with better statistical properties than former person fit indices. Three data tables and eight graphs are included. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310123

Detecting Aberrant Response Patterns in the Rasch Model

Rapport
87-3

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

J. Kogut

TM013696

Toegepaste Onderwijskunde

2

Vakgroep
Onderwijskundige Meetmethoden
en Data-analyse

Universiteit
Twente

Project Psychometrische Aspecten van Item Banking No.18

Colophon

Typing : L. Padberg
Cover design : M. Driessen, AV-section, University of Twente
Printed by : Central Reproduction Department, University of Twente

Detecting Aberrant Response Patterns in the Rasch Model

Jan Kogut

Abstract

In this paper, the detection of response patterns aberrant from the Rasch model is considered. For this purpose, a new person fit index, recently developed by Molenaar, and an iterative estimation procedure are used in a simulation study of Rasch model data mixed with aberrant data. Three kinds of aberrant response behavior are considered: guessing to complete the test, guessing in accordance with the three-parameter logistic model, and responding with different abilities on different subsets of items. The power in detecting such aberrants is evaluated in two cases: item difficulties known, and item difficulties estimated from data including aberrants. The results reveal that in the latter case, the estimates of the model parameters are biased and the power of the index, as a consequence, is reduced. It is shown that using an iterative procedure, the recovery of the power of the index to the level obtained by known item difficulties is achieved. Furthermore, dependent on the type of aberrance, a considerable reduction of the bias in the model parameters is possible. Finally, it is confirmed that this new index allows us to detect aberrant response patterns with better statistical properties than former person fit indices.

Introduction

In applications of the Rasch model (RM) it is often assumed that the data may contain response patterns from a minority of persons whose response behavior is aberrant. In such the case aberrants' response patterns should be detected and then treated separately because mostly the estimates of the aberrants' true ability are not appropriate, or at least, less reliable. Further, the removal of aberrants from the data might result in a reduction of the bias in the estimates of the model parameters.

The decision which person is of which behavior can only be taken on the base of his/her response pattern. In order to detect aberrant response patterns, various indices have been proposed (e.g., Drasgow, Levine & Williams, 1985; Molenaar & Hoijtink, 1987; Smith, 1985, 1986; Tatsuoka, 1984; Trabin & Weiss, 1983; Wright & Stone, 1979, chap.4, chap.7;). For a review of the indices and their application, see Kogut (1986). Usually, aberrant response patterns in the data are detected by comparing the value of a given person fit index with the 100*α-th percentile of its distribution under the RM. In such decision processes, the probability of misclassifying a RM-behaving person as aberrant will be equal to α.

Until recently, the indices in the RM were calculated conditionally on a fixed ability level. Now it is clear that such conditioning forces one to compare a given response pattern with all possible patterns of which many must have totally different ability estimates (Hoijtink, 1986; Molenaar & Hoijtink, 1987). In

contrast, conditioning on the total score in the RM limits the possible comparisons of a given pattern only to those for which this conditioning is appropriate (all these patterns have the same ability estimate). In addition, when conditioning on a fixed ability level, the distribution of patterns depends on the ability level. When conditioning on a total score, however, the distribution of patterns is independent of the ability; thus, a given pattern can be handled irrespective of the ability level it was obtained at.

So far, decisions about the fit of a given pattern to the model were made under the too optimistic supposition that the index distribution could be approximated by a normal distribution. Now, it is also evident that a more accurate approximation of the exact distribution of the index is needed. To realize these two recommendations, a new index based on conditioning on the total score, together with a more accurate approximation of its distribution have been proposed (Hoijtink, 1986; Molenaar & Hoijtink, 1987).

The object of this paper is to evaluate the power of Molenaar's new index in detecting persons of aberrant behavior. For this purpose, a simulation study was conducted. Three kinds of possibly aberrant response behaviors from the RM were considered: guessing to complete the test, guessing in accordance with the three-parameter logistic model, and responding with two different abilities on two different subsets of items. The detection of the aberrants was carried out in two different ways: 1) the generated item difficulties of the RM were known, for instance, from a

calibration study, and 2) the item difficulties were estimated from data containing some aberrants. In the latter case, the evaluation of the power of the Molenaar's index was carried out with the help of an iterative estimation procedure proposed in Kogut (1986). Furthermore, a comparison of the power of Molenaar's and Levine's index (Drasgow, Levine & Williams, 1985) was made.

Method

In this study, person fit to the RM is assessed using the previously mentioned Molenaar index (Hoijtink, 1986; Molenaar & Hoijtink, 1987), i.e.,

$$P(\underline{X} \mid r) = \frac{1}{\gamma_r} \prod_{i=1}^{k} \{\exp(-b_i)\}^{X_{vi}} ,$$

where b_i is the difficulty of item i, $r = \sum_{i=1}^{k} X_i$ is the total score for a pattern $\underline{X}=(X_1,\ldots,X_k)$ on a test with k items, and γ_r is the basic symmetric function of order r (Fischer, 1974). The value of the $P(\underline{X}|r)$ index for a given pattern is equal to the probability of the pattern in the RM given its total score (Fischer, 1974). This implies that if for a given pattern, the value of the $P(\underline{X}|r)$ index is very low, then this pattern is very improbable for a RM-behaving person. Therefore, if the index value is lower than a fixed percentile of the $P(\underline{X}|r)$ distribution for the RM, the proper decision is to consider the pattern as aberrant.

The fixed percentile for the distribution of $P(\underline{X}|r)$ index is rather a complicate function of the item difficulties because of

the nature of the γ_r's. If the item difficulties are known, the percentile can be calculated with the help of a complete enumeration of the (exact) distribution of the index. As this enumeration requires $\binom{k}{r}$ calculations for a given r, and 2^k-2 calculations for all r=1,...,k-1, it can be used only for small values of k. If there are many items, it is more convenient to use the 100*α-th percentile of an appropriate chi-square approximation to the index distribution (Molenaar & Hoijtink, 1987). Another possibility is to estimate the percentile from a sample distribution of RM patterns.

If the item difficulties are unknown and have to be estimated from data including aberrant response patterns, then a new impediment may arise. Due to the presence of aberrants, the estimates of the item difficulties usually are biased, and so are values of the $P(\underline{X}|r)$ index and the 100*α-th percentile. To remove the bias, an iterative strategy is proposed where each iteration consists of an approximation of the 100*α-th percentile, the decision which patterns are aberrant, and the actual removal of these patterns from the data, respectively.

Generation of Data

In order to evaluate the power of the $P(\underline{X}|r)$ index in detecting persons of an aberrant behavior, a simulation study was conducted. Both RM data and data containing aberrant response patterns were generated.

The RM data were composed of 2500 response patterns generated according to the RM for a test of 20 items. The item difficulties

and the distribution of ability were taken from cases already studied by Hoijtink (1986). The item difficulties were placed symmetrically around zero with more density in the neigbourhood of zero (namely: $\pm$0.11, $\pm$0.32, $\pm$0.54, $\pm$0.77, $\pm$1.02, $\pm$1.28, $\pm$1.58, $\pm$1.92, $\pm$2.35, $\pm$3.00). The distribution of ability was normal, N(0.0, 1.53). Such values for the item difficulties and this distribution of ability can be met when a test is specially designed for the group of homogeneous persons.

The aberrant response data consisted of 500 patterns generated to get a specific aberrance from the RM. Three kinds of possibly aberrant response behavior were considered: guessing to complete the test, guessing in accordance with the three-parameter logistic model, and responding with two different abilities by the same person on two different subsets of items.

Guessing to complete the test occurs when a person responds to some items at random, whereas his/her responses to the rest of the items are according to the RM. This deviation from RM response behavior is often observable with persons of low ability on the most difficult items. However, in tests with a time limit there might be no connection with the person's ability and/or the item difficulties. For simulation purposes, two extreme subset of items -the five most difficult and the five easiest- were selected from the item difficulties. Besides, in order to simulate aberrance of this kind, in particular for persons of low ability, three normal distributions of ability were selected: N(m, 1.53), where m=0.0, -1.0 and -2.0. On the selected subsets of items the responses were generated with a constant probability of a correct response equal

to 0.2, 0.25 and 0.5; in other words, guessing at random on multiple-choice items with five, four and two alternatives was simulated. On the rest of the items, responses according to the RM were generated. So, in all, eighteen data sets were generated, where each data set included 500 aberrant patterns of the type considered here.

Next, guessing in accordance with the three-parameter logistic (3PL) model was simulated. In applications of IRT models to data from multiple-choice tests, the 3PL model is supposed to handle guessing behavior on difficult items for persons of low ability more adequately. The only difference between the 3PL model and the RM is that the probability of the correct response for person v to item i, P_{vi}, now is a function of three item parameters (difficulty b_i, discrimination a_i, and pseudo-guessing parameter c_i). More precisely,

$$P_{vi}(\theta) = c_i + \frac{1-c_i}{1+\exp[-1.7a_i(\theta_v-b_i)]}, \quad v=1,\ldots,n;\ i=1,\ldots,k;$$

where $a_i>0$, $b_i \in R$ and $0 \leq c_i<1$ are the parameters characterizing item i in the 3PL model (in the RM we have $c_i=0$ and $a_i=1$ for all $i=1,\ldots,k$). In order to simulate a more involved type of guessing behavior, for all items the parameters of the 3PL model were intentionally fixed at the following values: all discrimination parameters were set equal to one, the difficulty parameters were set at the same level as in the RM, and the pseudo-guessing parameters at one of the values 0.2, 0.25 and 0.5. The response patterns were generated again using the above three different

distributions of ability: N(m, 1.53) with m=0.0,-1.0 and -2.0. So, nine data sets were generated, each consisting of 500 response patterns according to the 3PL model.

Finally, if a person responds with a varying ability to different items, his/her response behavior must be seen as aberrant from the RM as well. This aberrance is of frequent occurence for persons with cultural and educational retardation or with certain misconceptions. This phenomenon may also occurs if someone copies from a neighbour, or if certain item order by person interactions arise (e.g., a slow startup and sleepness). In this study, the case of a person displaying two distinct abilities on two different subsets of items was also simulated. As in the case of guessing to complete, the subsets of items were the five easiest and the five most difficult. The ability distribution of aberrants was the same as in the RM data, i.e., N(0.0, 1.53). On each of the subsets the abilities were lowered about 1 or 2 logits in comparison with the abilities on the rest of the items. Nevertheless, on both subsets the responses were generated according to the RM. So, four data sets, each consisting of 500 aberrant patterns, were used.

Approximation of the 100*α-th Percentile of the Index

To approximate the 100*α-th percentile of $P(\underline{X}|r)$, RM patterns were sampled because of a large number (20) of items considered. The approximation was carried out using the following three steps:

(1) RM patterns were sampled using estimates of the item difficulties and the person abilities to get at least 200 patterns for each total score r, r=1,...,19;

(2) the index values for the RM patterns sampled in (1) were calculated. Then, these values were collected and ordered per total score group (for each separate total score r, from the lowest, V_1, to the highest, V_{n_r}, index value);

(3) the 100*α-th percentile for the distributions of the index sampled in (2) were approximated by value of $V_{[\alpha * n_r + 1]}$ for each total score r.

The Power of the Index in Detecting Aberrants

For each generated data set, the value of the $P(\underline{X}|r)$ index was compared with the estimates of the 5%-th percentile. If the index value was lower than the estimate of this percentile, the observed pattern was classified as aberrant. This implies that the percentages of the RM-generated patterns misclassified as aberrant were expected to be about 5% , both for each total score separately and across total scores.

a) Item Difficulties Known from a Calibration Study

In applications of the RM we can possibly deal with cases where the item difficulties are known (for instance, from a careful calibration study). Accordingly, such a case was considered in this paper. To evaluate the power of the $P(\underline{X}|r)$ index in such the case, estimates of the item difficulties and the abilities were calculated from the RM data only by the conditional maximum likelihood method as implemented in the PML algorithm (Gustafsson, 1981). Further, the three steps to get an approximation of the 5%-th percentile of the index were carried out. Finally, the values of the $P(\underline{X}|r)$ index for the RM and the aberrant patterns were

calculated, and the decisions about aberrance were made with the help of the approximate percentiles.

b) Item Difficulties Estimated from Data Including Aberrants

On the other hand, in applications of the RM we have to deal with cases where the item difficulties are unknown and have to be estimated from data containing aberrant response patterns. Therefore in this study the item difficulties and abilities were also estimated from RM data mixed with data from aberrants. Having these estimates, the power of the index was evaluated as in the previous case. However, here the whole procedure was carried out iteratively. This means that after the detection of aberrant patterns, they were removed from the data and the procedure was repeated until new aberrant patterns were no longer found.

In order to compare the power of the $P(\underline{X}|r)$ index (conditioning on the total score) with the power of the former indices conditional on a fixed ability level, the ZL_0 index was used (Drasgow, Levine & Williams, 1985):

$$ZL_0 = \frac{L_0 - E\ (L_0|\hat{\theta})}{Var^{1/2}\ (L_0|\hat{\theta})},$$

where

$$L_0 = \log P(\underline{X}|\hat{\theta}) = \log\{\prod_{i=1}^{k} P_{vi}(\hat{\theta})^{X_{vi}}[1-P_{vi}(\hat{\theta})]^{1-X_{vi}}\},$$

and $E(L_0|\hat{\theta})$, $Var(L_0|\hat{\theta})$ are the conditional expected value and variance of L_0, respectively. Note that the ZL_0 index is the

standardized version of L_0, the origin of $P(\underline{X}|r)$. Working with the ZL_0 index, the standard normal approximation was used if a decision about aberrance had to be made. When the obtained value of ZL_0 index was smaller than -1.96, the observed pattern was classified as aberrant.

Results

Guessing to Complete the Test

The results for the mean power of the $P(\underline{X}|r)$ index in detecting aberrance, are for the case of known item difficulties, presented in Table 1, Colomn 3. These results are indicative of the percentages of aberrants detected correctly with the $P(\underline{X}|r)$ index over the three groups of aberrants with a normal distribution of ability. As is clear from Table 1, the mean power of the $P(\underline{X}|r)$ index depends to a high degree on the probability of guessing, on the mean ability of the aberrants, and on the item difficulties of the guessed items.

Insert Table 1 about here

In the case of the five most difficult items, the mean power to detect guessing to complete increases with the probability of guessing and decreases with the mean ability of the guessers. In

contrast, for the five easiest items, the mean power decreases with the probability of guessing and increases with the mean ability of the guessers.

Actually, such results are to be expected if we notice the inconsistency between the aberrant and RM patterns. For instance, guessing at random on five items with the probability of guessing equal to 0.2 results in 0, 1 or 2 correct responses with the cumulative binomial probability about 0.94. The same cumulative probability with the guessing constant 0.5 will be obtained for 1, 2, 3 and 4 correct responses. For guessers, these predictions are independent of the ability and the item difficulties of the items the person guesses on. However, for a RM-behaving person, the probability of the correct responses depends to a high degree on his/her ability and on the item difficulties. To clear this question up, let us consider the item characteristic curves (ICC's) of the RM at a fixed ability level. In the case of five difficult items it holds that the lower the ability of a RM-behaving person, the higher the probability of incorrect responses on all of these items. Therefore, the inconsistency between possibly random and RM responses will increase with the probability of guessing but decrease with the ability. In turn, with increasing inconsistency, the power to detect aberrance will rise. This dependency, i.e, the power as a function of ability for the probability of guessing equal to 0.25, is ilustrated in Figure 3 (see below). In the case of five easy items, even a RM-behaving person of relatively low ability should response correctly on some of these items. Therefore, the inconsistency and the power will decrease with the

probability of guessing but increase with the ability of person. For this case and for the probability of guessing equal to 0.25, the power as a function of ability is illustrated in Figure 1. If such a function is known, conclusions about the mean power over the group (as the values presented in Table 1) can be drawn.

The comparison between the mean power of the $P(\underline{X}|r)$ and ZL_0 indices (Colomns 3 and 4 of Table 1, respectively) clearly indicates the superiority of the $P(\underline{X}|r)$ index, irrespective of the conditions under which the power is evaluated. This confirms that conditioning on the total score and using the given approximation of the $100*\alpha$-th percentile of the index results in more effective detection of aberrants. In addition, when using the $P(\underline{X}|r)$ index, unlike ZL_0, it is also possible to obtain a constant Type I error. Namely, for the $P(\underline{X}|r)$ index, the type I error was 4.88% over the group of all RM patterns and about 5%, with larger random deviations, for each total score group separately. For ZL_0, this error was 2.20% on the average but 0.0% for the extreme and about 3.5% for the middle total score group. Thus, the ZL_0 index is more conservative in detecting aberrants.

The power of the $P(\underline{X}|r)$ index for the case of the item difficulties estimated from data containing aberrants - thus when the iterative procedure was used - is given in Figure 1. Here the ability distribution of aberrants was the same as for the RM data, i.e., normal N(0.0, 1.53). Besides, aberrants responded at random with the probability of guesssing equal to 0.25 on the subset of

the five easiest items. It should be reminded that in the case of known item difficulties, the mean power was 72.0% (see Table 1) and the power was a rapidly increasing function of ability (see Figure 1, Graph •).

Insert Figure 1 about here

Note that at a broad range of high abilities the power to detect this aberrance was nearly 100% . When using estimates of the item difficulties based on the data containing all aberrants the power is significantly lower. Nevertheless, after the first two iterations a considerable improvement of the detection of aberrants was obtained. Note also that the power converged almost monotonically to the power for the case of known item difficulties; however, the type I error after the third iteration was a little larger (5.94%) than for the case of known item difficulties (4.88%).

The bias in the estimates of the item difficulties due to the presence of about 20% aberrants is, for the four subsequent iterations, given in Figure 2. At the first iteration,

---

Insert Figure 2 about here

---

thus before any aberrants were removed, the PML estimates of the item difficulties had been systematically biased. From Figure 2 the large overestimation of difficulties of the items the aberrants guessed on and the underestimation of the higher item difficulties is obvious. When the item difficulty approached the extremes of the difficulty scale, over- and underestimation increased (to about 13 and about 3 times their standard errors, respectively). Using the iterative procedure, a large part of the bias was reduced. A very significant reduction of bias in the estimates of item difficulties was obtained after the first two iterations. Although subsequent iterations still reduced the bias for the guessed items, at the same time, they introduced an another bias in the opposite direction (to be seen at the right extreme of the difficulty scale). This new bias is due to the removal of the RM patterns that were misclassified as aberrant (Type I error). Besides, it might be due to the aberrant patterns still remaining in the data because of a too small inconsistency to be detected by the $P(\underline{X}|r)$ index (lack of power of the index).

On the other hand, the presence of about 20% aberrants affects the estimates of the abilities in a similar way. The low abilities were overestimated whereas the high abilities were underestimated. Here, however, the bias changed monotonically and was much lower than the standard errors of the ability estimates (to about 39% and about 22% of the standard errors at the extremes of the ability scale). Using the iterative procedure a reduction of the bias in the estimates of abilities was possible as well. The maximal reduction of bias was obtained after two iterations. In the

subsequent iterations the bias increased.

Let us now consider the event of aberrants of low ability guessing at random on the five most difficult items. Here the ability distribution of aberrants was normal N(-2.0, 1.53), and as in the previous case, the probability of guesssing was equal to 0.25. It should be reminded that in the case of known item difficulties, the mean power for the $P(\underline{X}|r)$ index was 37.6% (see Table 1), and the power was a decreasing function of ability (see Figure 3, Graph •). From Figure 3 it is also evident

---

Insert Figure 3 about here

---

that detecting this kind of aberrance with 100% certainly was not possible, even not for persons of a very low ability. On the other hand, using estimates of the item difficulties obtained from the data containing all aberrants did not reduce the power of the index significantly. These results are thus in contrast with those for the five easiest items. However, the former level of the power from the case of known item difficulties was almost reached when the iterative procedure was used, after the second iteration.

The bias in the estimates of the item difficulties in question was in general smaller (see Figure 4).

Insert Figure 4 about here

Almost complete reduction of the bias on the guessed items was observed after the second iteration. Also, the optimal reduction of bias over all items was obtained after two iterations. The next iterations introduced another bias of the opposite direction, in particular for the items with difficulties at the left extreme of the scale (as can be seen in Figure 4).

Guessing in Accordance with the 3PL Model

For the case of known item difficulties, the mean power of the $P(\underline{X}|r)$ index to detect guessing in accordance with the 3PL model is shown in Table 2. It can be seen that the mean power

Insert Table 2 about here

of the $P(\underline{X}|r)$ index increased with the pseudo-guessing parameter and decreased with the mean ability. These results can be expected if we notice the inconsistency between patterns obtained in the 3PL model and in the RM. For this purpose, consider the ICC's for the 3PL and the RM for the ability level fixed at θ. The differences

between these ICC's increase with the difficulty of the items, particularly for items with a difficulty on the right side of θ. For these items, a 3PL pattern tends to contain more correct responses than a RM pattern. If a person's ability is much below the difficulty of the easiest item, the 3PL pattern still might contain a certain number of correct responses placed at random over the items. So, detecting this aberrance for a person of very low ability will be easy, whereas for a person of very high ability it may be almost impossible (see Figure 5 for this dependency). Obviously, with increasing pseudo-guessing parameter values, the differences between the 3PL and RM patterns tend to be larger; hence, the detection of this aberrance should be improved.

The power of the $P(\underline{X}|r)$ index for item difficulties estimated from data including aberrant patterns is shown in Figure 5. Here the iterative procedure was applied as well. The ability distribution of the aberrants was N(-2.0, 1.53) and the pseudo-guessing parameters were set equal to 0.25 (see Table 2 for the mean power in this case). As it can be seen from Figure 5, a considerable improvement of the power in detecting aberrants was

---

Insert Figure 5 about here

---

obtained after the first iteration only. These results correspond to the case of known item difficulties. The next iterations showed

a little increase in power but at the cost of increasing the Type I error (about 5.98% after the third iteration).

The bias in the estimates of the item difficulties due to the presence of the aberrants is shown in Figure 6.

Insert Figure 6 about here

The presence of 500 aberrant patterns resulted in overestimation of the easy item difficulties and underestimation of the difficult items. This bias changed uniformly and was maximal at the extremes of the difficulty scale (about 5 times the standard errors). After the first iteration, most items had difficulty estimates witnin their standard errors. Subsequent iterations introduced a bias in the opposite direction; however, this was only observable for a few items of extreme difficulty.

Responding with Two Different Abilities on Different Subsets of Items

The mean power of the $P(\underline{X}|r)$ index to detect aberrance of this kind, for the case of known item dificulties, is given in Table 3.

Insert Table 3 about here

Notice that in this case, for a mean ability of guessers equal to 0.0, the $P(\underline{X}|r)$ index had much less power than in the case of guessing to complete the test, implying that this type of aberrance is more difficult to detect by the $P(\underline{X}|r)$ index. In order to explain this, let us consider the ICC's in the RM, in particular for these items on which the aberrance occurred. For the case of five easy items, if an aberrant's ability (i.e., the one on the rest of items) corresponds to the difficulty of the items in question, then the larger the difference between the two abilities, the larger the modifications in the response pattern compared with the expected RM pattern. This is why such an aberrant person can be detected. If an aberrant's ability differs much from the difficulties of the five easy items, then less modifications in the pattern can be expected; thus, such a person will be more difficult to detect (this dependency can be seen in Figure 7; however, for very high abilities only). Now, let us consider the subset of the five difficult items. If an aberrant's ability is below the difficulties of these items, then the response pattern should contain fewer inconsistencies. Therefore, the percentages of detected aberrants will be below supposed 5% (see Table 3).

The power of the $P(\underline{X}|r)$ index for item difficulties estimated from data including aberrant patterns is shown in Figure 7. Here the ability distribution of aberrants was N(0.0, 1.53) and the aberrants had a ability lowered by 2.0 logits on the five easiest items compared to the ability on the rest of the items. As it is shown in Figure 7,

Insert Figure 7 about here

a small but uniform improvement of the power to detect aberrants correctly can be observed for the first two iterations. Here it also seems that the power is converging monotonically to the one for the case of known item difficulties.

The bias in the estimates of the item difficulties due to the presence of the aberrants is shown in Figure 8.

Insert Figure 8 about here

Again over- and underestimation of the item difficulties were observed at the first iteration, i.e, when all aberrant patterns were present in the data. The overestimation occured on the items on which there was aberrance, and underestimation on the rest of items. It is remarkable that on both subsets, the bias on the items was more uniform than the one in the case of guessing to complete. When using the iterative procedure the optimal reduction of bias seemed to be obtained after the second or third iteration, but this reduction was not fully satisfactory.

It can be expected that other indices, specially developed for

the aim to detect this type of aberrance (e.g., the unweighted between fit index by Smith, 1985) will be more efficient. The cost of this will then, of course, be a reduced power to detect other types of aberrance (Smith, 1985,1986).

Discussion

The results of this simulation study show that in appplications of the Rasch model to test data the $P(\underline{X}|r)$ index is more successful in detecting aberrant reponse patterns than the formerly more popular ZL_0 index. Namely, more power in the detection and a preassigned Type I error is obtained. This confirms the expected advantages with respect to conditioning on the total score (instead of on a fixed ability) and the use of a more accurate approximation to the exact distribution of the index (instead of a normal approximation).

Further, the dependency of the power of the index on ability has been seen to vary according to the kind of aberrance. For guessing on easy items to complete the test, the power is a rapidly increasing function of ability, but for guessing on difficult items as well as for guessing in accordance with the 3PL model it is a decreasing function. However, for responding with two different abilities by the same person on two different subsets of items no clear dependency could be observed.

If item difficulties are unknown and have to be estimated from data containing aberrant response patterns, then, in general, the power of the $P(\underline{X}|r)$ index is reduced. The results show that using

the iterative procedure, the previous level of the power is recovered within a few iterations. This is, however, at the cost of a small increment in the Type I error. For these reasons, the use of the iterative procedure can be recommended if the detection of aberrants is a problem of interest. In such cases only a few iterations should be used.

Finally, due to the presence of aberrant patterns in the data, estimates of the item difficulties might be very biased dependent on the kind of aberrance. If the aberrance occurs on the whole test, the use of two or three iterations of the procedure generally results in a satisfactory reduction of the bias. If the aberrance occurs on a few items only, the results are questionable and not so satisfactory because of remaining bias. This final bias is due to the removal of the RM patterns misclassified as aberrant. Besides, it might be also due to the presence of aberrant patterns left in the data that could not be detected by the index. So, if there is no other way to estimate the item difficulties, the use of the iterative procedure with only a few iterations is recommended. However, one should take into account that some bias may remain in the estimates.

References

Birenbaum, M. (1985). Comparing the effectivenees of several IRT based appropiateness measures in detecting unusual response paterrns. Educational and Psychological Measurement, 45, 523-534.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropiateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Fischer, G.H. (1974). Einführung in die Theorie psychologicher Tests. Bern:Huber.

Gustafsson, J.E. (1981). PML: a computer program for conditional estimation and testing in the Rasch model for dichotomous items. Reports from the institute of education, University of Goteborg, no. 85.

Harnisch, D.L., & Tatsuoka, K.K. (1983). A comparision of appropiateness indices based on item response theory. In R.K. Hambleton (Ed.), Applications of item response theory. Vancover, B.C.: Educational Reseach Institute of British Colombia.

Hoijtink, H. (1986a). Detecting aberrant response patterns in the unidimensional scaling model of Rasch. Heymans Bulletin, HB-86-792-SW. Groningen: R.U.Groningen.

Hoijtink, H. (1986b). Rasch schaal constructie en person fit. (Technical Report). Groningen: Psychologische Instituten, R.U.Groningen.

Kogut, J. (1986). Review of IRT-based indices for detecting and diagnosing aberrant response patterns. Rapport 86-4. The Netherlands: Enschede, University of Twente, Department of Education.

Molenaar, I.W & Hoijtink, H. (1987). The many null distributions of person fit indices. Heymans Bulletin, HB-87-846-EX. Groningen: R.U.Groningen.

Rudner, L.M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Smith, R.M. (1985). A comparision of Rasch person analysis and robust estimators. Educational and Psychological Measurement, 45, 433-444.

Smith, R.M. (1986). Person fit in the Rasch model. Educational and Psychological Measurement, 45, 433-444.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. Journal of Educational Statistics, 10, 55-73.

Trabin, E.T., & Weiss, D.J. (1983). The person response curve: fit of individuals to item response theory models. In D.J. Weiss (Ed.) New horizonts in testing: latent trait theory and computarized adaptive testing. New York: Academic Press.

Wright, B.D., & Stone, M.H. (1979). Best test design. Chicago: MESA Press.

Author's Note

The author is indebted to Ron J.H. Engelen and Wim J. van der Linden for their comments on an earlier version of the paper. The content of the paper is however fully his own responsibility.

Table 1

Mean power of $P(\underline{X}|r)$ and ZL_0 in detecting guessing to complete the test

Mean Ability of Guessers*	Probability of Guessing	Power (in %) $P(\underline{X}\|r)$	ZL_0
	On Five Most Difficult Items		
0.0	0.20	13.6	8.0
	0.25	16.8	11.6
	0.50	42.0	33.4
-1.0	0.20	21.0	12.6
	0.25	25.6	15.8
	0.50	55.6	47.2
-2.0	0.20	31.6	18.8
	0.25	37.6	23.6
	0.50	69.2	60.8
	On Five Most Easy Items		
0.0	0.20	76.2	63.9
	0.25	72.0	60.8
	0.50	48.0	41.2
-1.0	0.20	57.6	42.6
	0.25	52.6	39.2
	0.50	30.8	23.0
-2.0	0.20	38.2	22.8
	0.25	33.2	21.0
	0.50	17.2	11.4

* Ability of guessers is N(m, 1.53), where m=0.0, -1.0 and -2.0.

Table 2

Mean power of $P(\underline{X}|r)$ in detecting guessing in accordance with 3PL model

Mean Ability of Guessers*	Pseudo-guessing in 3PL	Power of $P(\underline{X}\|r)$ (in %)
0.0	0.20	17.9
	0.25	22.8
	0.50	32.6
-1.0	0.20	27.9
	0.25	31.3
	0.50	45.5
-2.0	0.20	39.0
	0.25	45.6
	0.50	62.5

* Ability of guessers is N(m, 1.53), where m=0.0, -1.0 and -2.0.

Table 3

Mean power of $P(\underline{X}|r)$ in detecting
two different abilities on different subsets of items

Difference in Abilities*	Power of $P(\underline{X}\|r)$ (in %)
On Five Most Difficult Items	
-1.0	1.8
-2.0	1.0
On Five Most Easy Items	
-1.0	13.8
-2.0	33.0

*Ability on the five items subset is lower than ability on the rest of items; (ability of guessers is N(0.0, 1.53)).

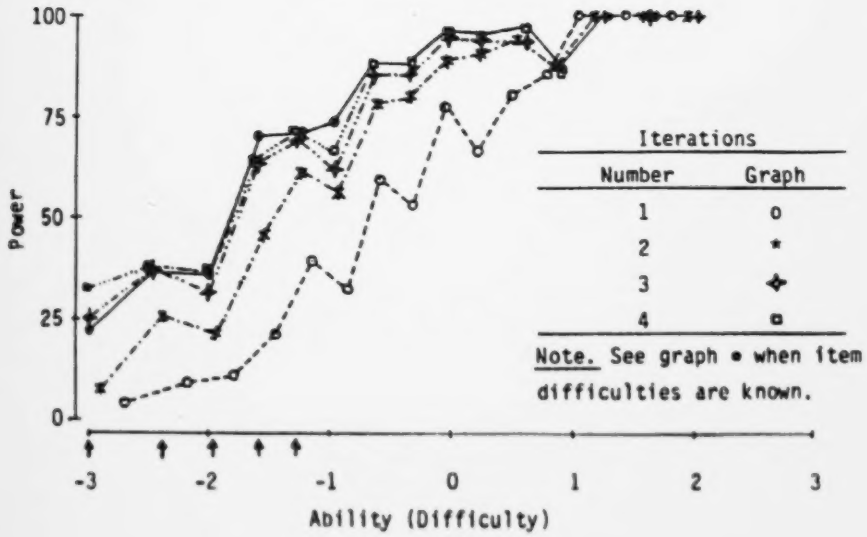

Figure 1. Power of $P(\underline{X}|r)$ in detecting guessing to complete on the five most easy items (probability of guessing is 0.25, ability of guessers is N(0.0, 1.53)).
Note. Difficulties of guessed items are marked with †.

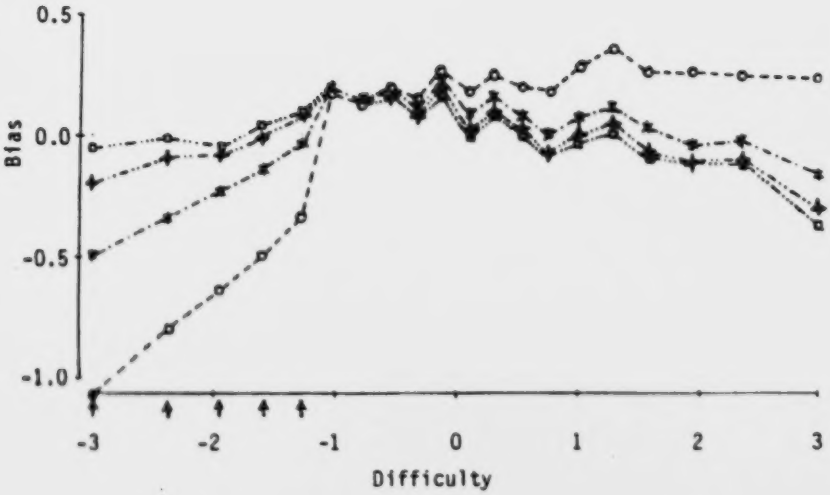

Figure 2. Bias in estimates of item difficulties by guessing to complete on the five most easy items (probability of guessing is 0.25, ability of guessers is N(0.0, 1.53)). (See Figure 1 for explanation of symbols).

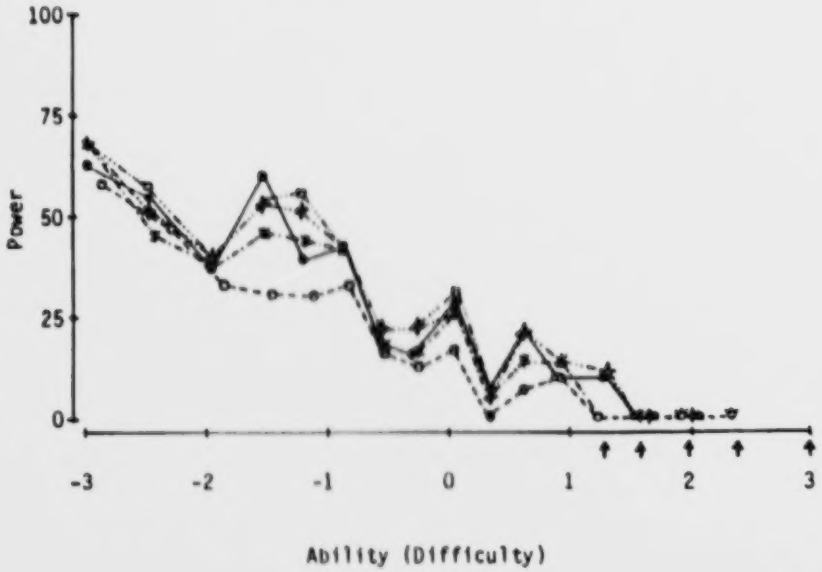

Figure 3. Power of $P(\underline{X}|r)$ in detecting guessing to complete on the five most difficult items (probability of guessing is 0.25, ability of guessers is N(-2.0, 1.53)). (See Figure 1 for explanation of symbols).

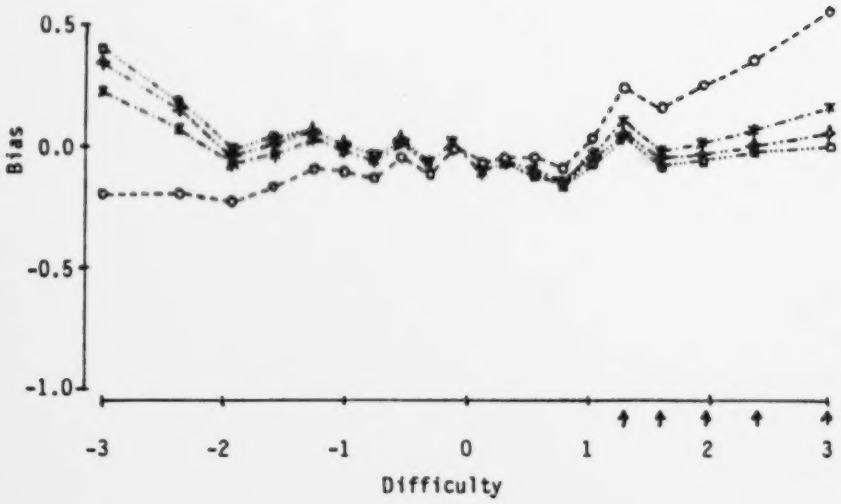

Figure 4. Bias in estimates of item difficulties by guessing to complete on the five most difficult items (probability of guessing is 0.25, ability of guessers is N(-2.0, 1.53)). (See Figure 1 for explanation of symbols).

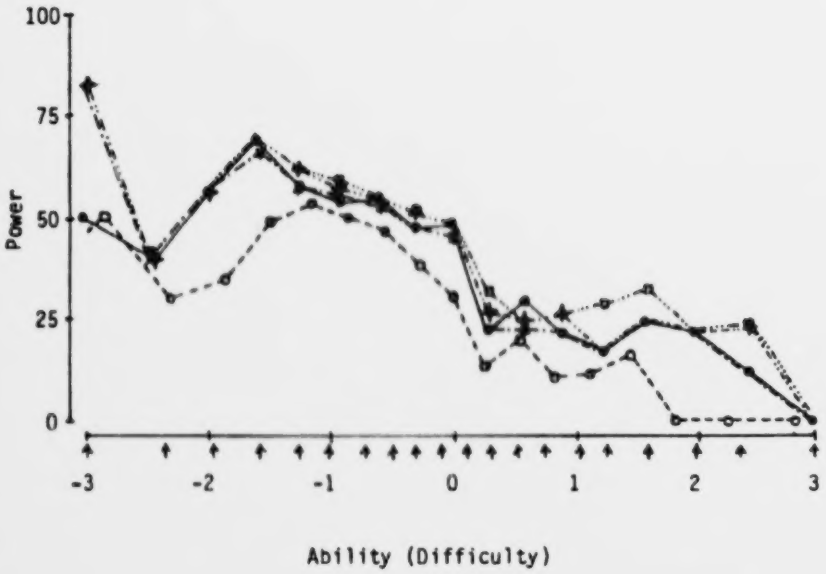

Figure 5. Power of $P(\underline{X}|r)$ in detecting guessing in accordance with the 3PL model (pseudo-guessing parameters are 0.25, ability of guessers is N(-2.0, 1.53)).
(See Figure 1 for explanation of symbols).

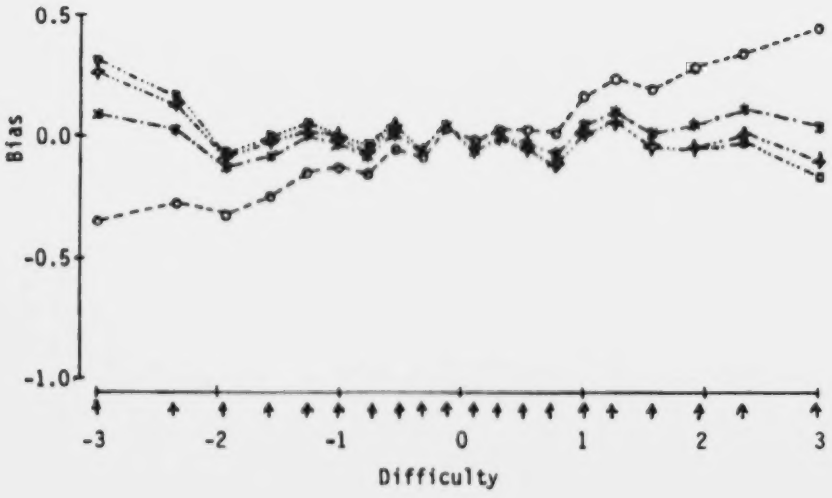

Figure 6. Bias in estimates of item difficulties by guessing in accordance with the 3PL model (pseudo-guessing parameters are 0.25, ability of guessers is N(-2.0, 1.53)). (See Figure 1 for explanation of symbols).

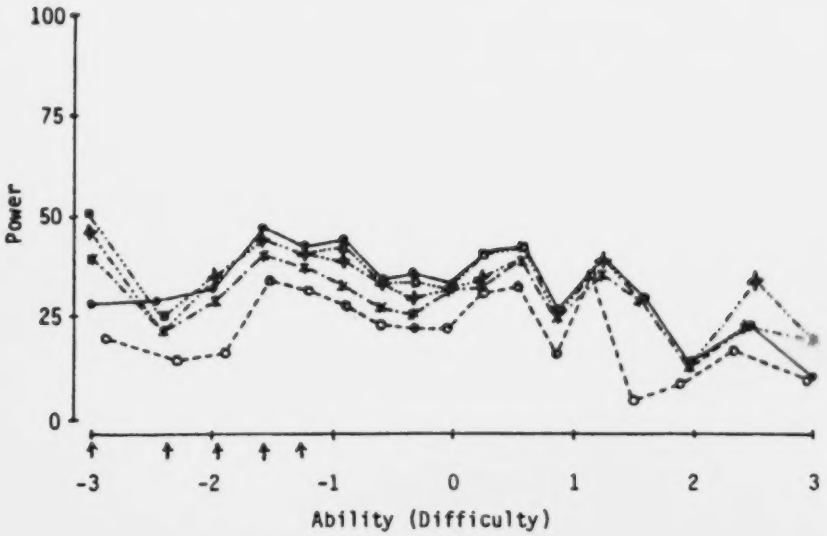

Figure 7. Power of P($\underline{X}$|r) in detecting two different abilities on two different subsets of items (ability on the five most easy items is 2.0 lower, ability of aberrants is N(0.0, 1.53)).
(See Figure 1 for explanation of symbols).

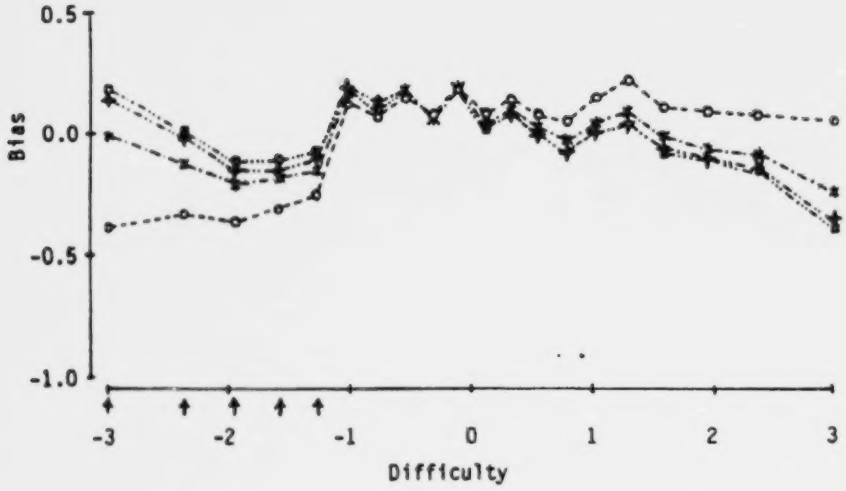

Figure 8. Bias in estimates of item difficulties by two different abilities on different subsets of items (ability on the five most easy items is 2.0 lower, ability of aberrants is N(0.0, 1.53)).
(See Figure 1 for explanation of symbols).

Titels van Recente Rapporten

R-86-1 C. Sluijter, Het bijwerken van een item bank

R-86-2 D.L. Knol, Inventarisatie van automatische itemselectie procedures voor Raschschalen

R-86-3 G.R. Buning, T.J.H.M. Eggen, H. Kelderman & W.J. van der Linden (red.), Het gebruik van het Raschmodel voor een decentraal toetsservicesysteem

R-86-4 J. Kogut, Review of IRT-based indices for detecting and diagnosing aberrant response patterns

R-86-5 D.L. Knol, Een overzicht van meerdimensionale itemresponsmodellen

R-87-1 J. Adema, Testconstructie met klassieke item- en testparameters

R-87-2 D.L. Knol, Het verband tussen itemresponstheorie en factoranalyse voor dichotome items

R-87-3 J. Kogut, Detecting aberrant response patterns in the Rasch model

Rapporten in deze reeks kunnen tegen kostprijs besteld worden bij: Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, Postbus 217, 7500 AE Enschede, tel. 053-893588

Titles of Recent Research Reports

RR-86-1 W.J. van der Linden, The use of test scores for classification decisions with threshold utility

RR-86-2 H. Kelderman, Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups

RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by zero-one programming

RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one programming approach to Gulliksen's matched random subtests method

RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features

RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes approach to item banking

RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of parallel tests by zero-one programming

RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models for paired comparisons with ties

RR-86-9 H. Kelderman, Common item equating using the loglinear Rasch model

RR-86-10 W.J. van der Linden, & M.A. Zwarts, Some procedures for computerized ability testing

RR-87-1 R. Engelen, Semiparametric estimation in the Rasch model

RR-87-2 W.J. van der Linden (Ed.), IRT-based test construction

RR-87-3 R. Engelen, P. Thommassen & W. Vervaat, Ignatov's theorem: A new and short proof

Research Reports can be obtained at costs from Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

Een publicatie
van de faculteit der
Toegepast Onderwijskunde
van de Universiteit Twente,
Postbus 217,
7500 AE Enschede

Toegepaste Onderwijskunde